Εργασία στα Γραμμικά Μοντέλα

Κωνσταντίνος Παπανίκος (1112201600167)
Κωνσταντίνα Τύραλη (1112201600230)
Δημήτρης Φούντας (1112201600236)

23 Ιουνίου 2020

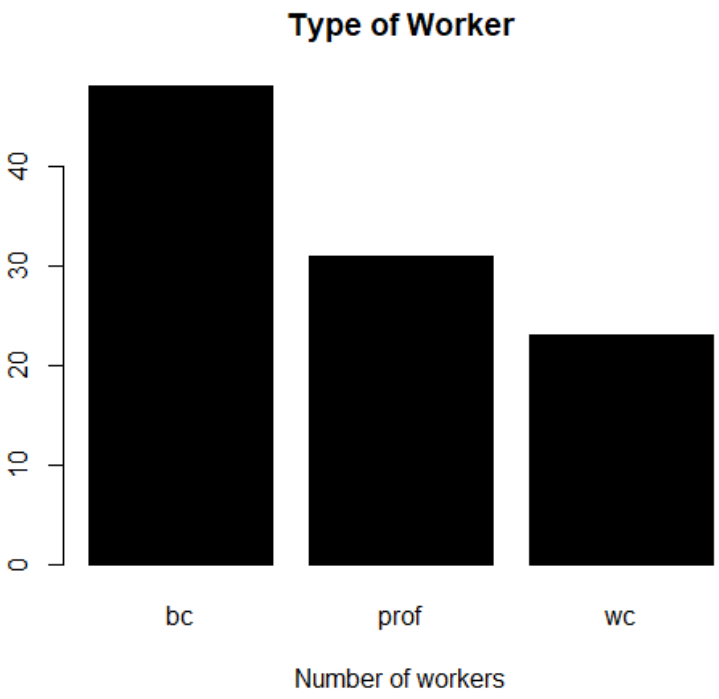
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

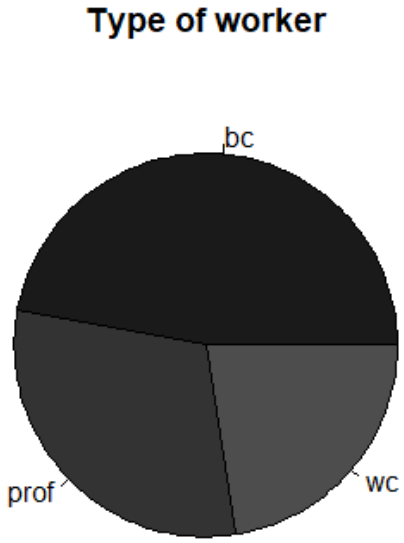
ΠΕΡΙΛΗΨΗ:

Αρχικά παρατηρούμε ότι το δείγμα μας όσον αφορά τον τύπο εργαζομένων δεν είναι ιδιαίτερα ομοιόμορφο, έχουμε 48 χειρονάκτες, 31 διοικητικούς και μόλις 23 επαγγελματίες γραφείου. Εξίσου εύκολα προσέχουμε πως το μέσο *prestige* του δείγματος μας είναι 46 στην κλίμακα που θέσπισαν οι $Pineo - Porter$ με την ψυχοκοινωνική μέθοδο που ανέπτυξαν κατά τη δεκαετία του 1960, η μέση εκπαίδευση είναι 10.54 χρόνια και το μέσο ποσοστό γυναικών ανά επάγγελμα είναι 13,6%. Βλέπουμε επίσης ότι υπάρχουν επαγγέλματα στο δείγμα μας τα οποία απαιτούν βασική μόρφωση 6,38 χρόνων ενώ υπάρχουν και επαγγέλματα που απαιτούν κοντά στα 16 χρόνια εκπαίδευσης. Παρομοίως, παρατηρούμε ότι παρ' όλο που το ποσοστό γυναικών ανά επάγγελμα είναι κατά μέσο όρο 13%, γίνεται εμφανές στα στοιχεία μας ότι υπάρχει τουλάχιστον ένα 'γυναικοκρατούμενο' επάγγελμα, έχοντας ποσοστά γυναικών εργαζομένων στο 97,5% (γραμματείς), ενώ υπάρχουν περισσότερα άπό ένα επαγγέλματα στα οποία εργάζονται αποκλειστικά άντρες (πυροσβέστες, μηχανικοί τρένων κλπ). Παρατηρούμε επίσης ότι η εργασία με το ελάχιστο εισόδημα είναι οι *babysitters* με αμοιβή 611$ ενώ το μεγαλύτερο εισόδημα του δείγματός μας βρίσκεται στο επάγγελμα του γενικού διευθυντή και ανέρχεται στα 25879$. Τέλος, βλέπουμε ότι το μικρότερο *prestige* (14.8) βρέθηκε στο επάγγελμα των *newsboys* ενώ το μεγαλύτερο (87.2) βρέθηκε στο επάγγελμα *physicians*. Στην συνέχεια παρατηρούμε ότι συνήθως όσο μεγαλύτερη είναι η εκπαίδευση ενός ατόμου (σε έτη), τόσο μεγαλύτερο και το *prestige* της εργασίας του. Όμως όταν συγκρίναμε την κοινωνικοοικονομική θέση με το εισόδημα παρατηρήσαμε κάτι εξαιρετικά ενδιαφέρον, όταν το *prestige* είναι μικρότερο του 80, όσο μεγαλώνει αυτό μεγαλώνει ανάλογα και το εισόδημα, αλλά όταν το *prestige* ξεπερνάει το 80 το εισόδημα του επαγγελματία αυξάνεται με πολύ μεγαλύτερους ρυθμούς από πριν, με περαιτέρω αύξηση του *prestige*. Κάταλήξαμε επίσης στο συμπέρασμα ότι η κοινωνικοοικονομική θέση που αντιστοιχεί σε ένα επάγγελμα εξαρτάται άμεσα από το εισόδημα και την εκπαίδευση, μάλιστα αν υπάρχει γνώση και των δύο, μπορεί να γίνει μια αρκετά καλή πρόβλεψη για το πόσο *prestige* έχει το συγκεκριμένο επάγγελμα. Για παράδειγμα, ένας εργαζόμενος με 10 χρόνια εκπαίδευσης και 4000$ εισόδημα προβλέπεται να έχει περίπου 42.87 μονάδες στην κλίμακα του *prestige*. Στην συνέχεια εξετάζουμε το *prestige* σε σχέση με το εισόδημα και την εκπαίδευση ταυτόχρονα αντί για να τα εξετάσουμε ξεχωριστά και βλέπουμε ότι το μοντέλο αυτό μας δίνει πιο αξιόπιστα αποτελέσματα απ' ότι όταν τα διερευνήσαμε ξεχωριστά. Τέλος, χρειάστηκε έλεγχος με ανάλυση διασποράς για το εισόδημα σε σχέση με το αν το ποσοστό γυναικών εν ενεργεία επαγγελματιών είναι πάνω ή κάτω από 50% και το είδος του επαγγέλματος. Ουσιαστικά αυτό έγινε για να εξετάσουμε αν το είδος του επαγγέλματος και το ποσοστό των γυναικών επηρεάζουν τον αναμενόμενο μισθό. Έτσι δημιουργούνται 6 διαφορετικές περιπτώσεις:
α)πάνω από 50% γυναίκες και χειρώνακτες → αναμενόμενος μισθός: 3.200$
β)πάνω από 50% γυναίκες και επαγγελματίες, διοικητικοί και τεχνικοί → αναμενόμενος μισθός: 6.000$
γ)πάνω από 50% γυναίκες και επαγγελματίες γραφείου → αναμενόμενος μισθός: 4.000$
δ)κάτω από 50% γυναίκες και χειρώνακτες → αναμενόμενος μισθός 6.000$
ε)κάτω από 50% γυναίκες και επαγγελματίες, διοικητικοί και τεχνικοί → αμενόμενος μισθός: 11.800$
στ)κάτω από 50% γυναίκες και επαγγελματίες γραφείου → αναμενόμενος μισθός 6.500$

1ο Ερώτημα:

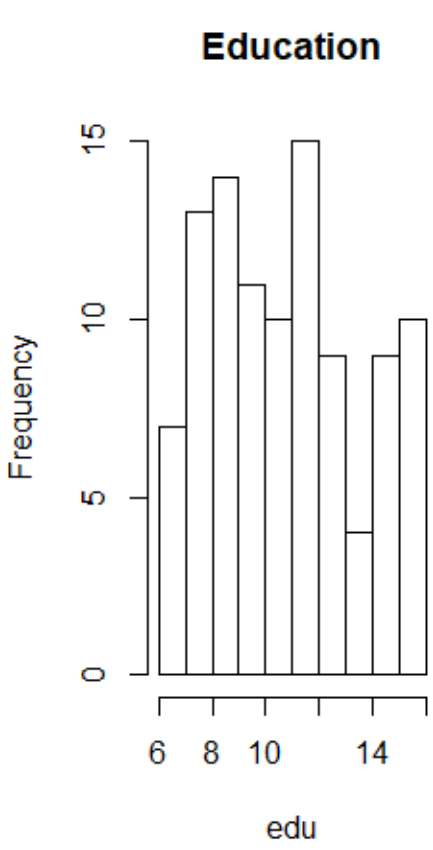
Για αρχή, χωρίζουμε τις μεταβλητές σε δύο κατηγορίες: *qualitative* και *quantitative*, καθώς η κάθε κατηγορία απαιτεί διαφορετική διαχείριση. Η μεταβλητή *type* δεν μπορεί να ποσοτικοποιηθεί, καθώς πρόκειται για τα τρία επίπεδα επαγγελμάτων, άρα ανήκει στην πρώτη περίπτωση. Έτσι, κάνουμε περιγραφική στατιστική και βρίσκουμε τον αριθμό των εργαζομένων από κάθε επίπεδο: 48 άτομα χειρώνακτες (*bc*), 31 επαγγελματίες, διοικητικοί και τεχνικοί (*prof*) και 23 επαγγελματίες γραφείου (*wc*). Έχουμε τα γραφήματα *barplot* και *pie*:

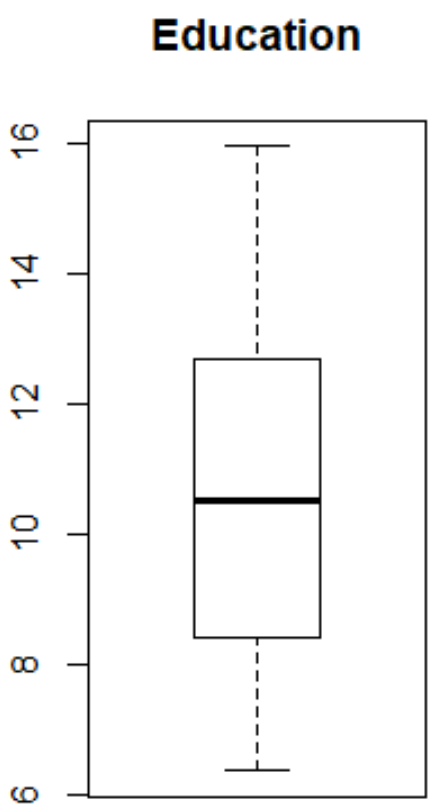


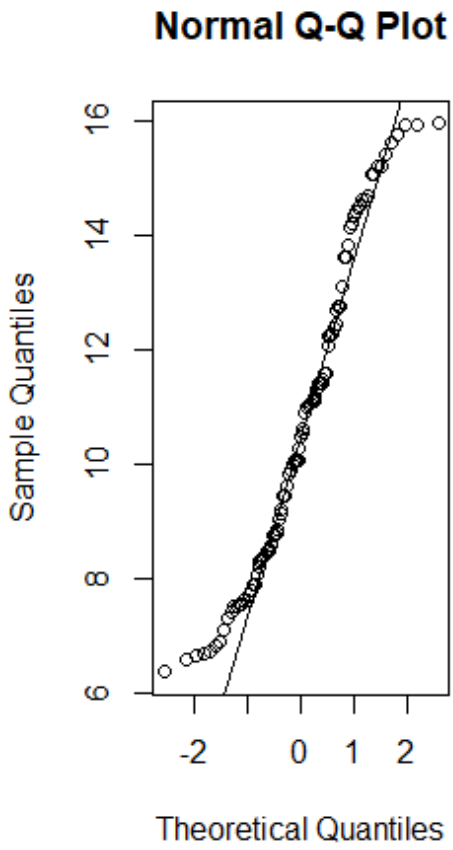


Τώρα, θέλουμε να κάνουμε περιγραφική στατιστική στα υπόλοιπα δεδομένα. Πρώτα θα ελέγξουμε ποιες από τις μεταβλητές μπορεί να ακολουθούν κανονική κατανομή μέσω ιστογράμματος *boxplot* και $q-qplot$.
Έχουμε λοιπόν:

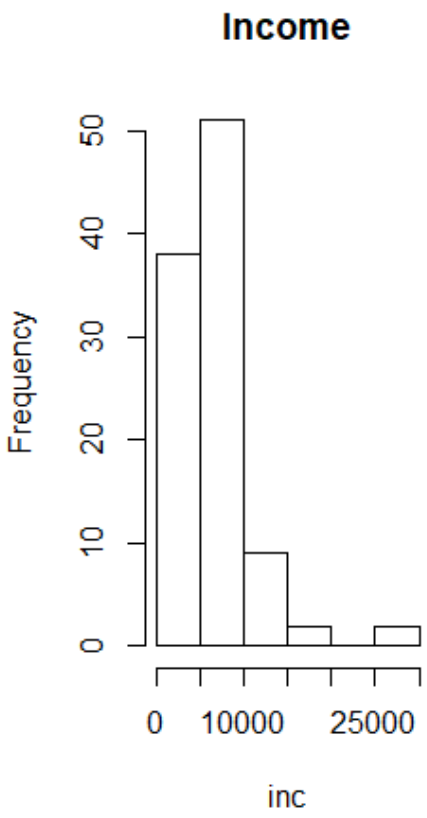
Για την εκπαίδευση:

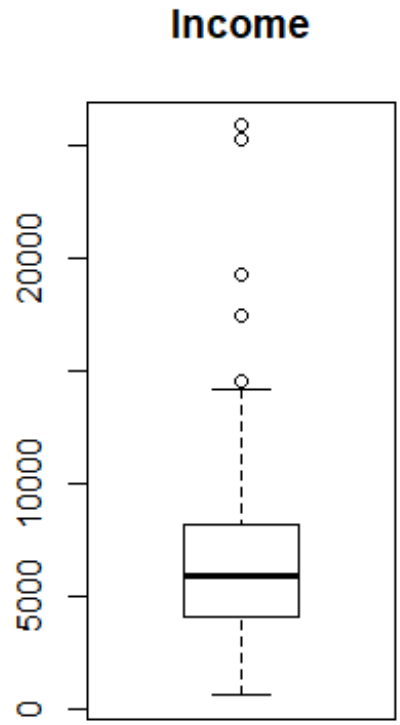


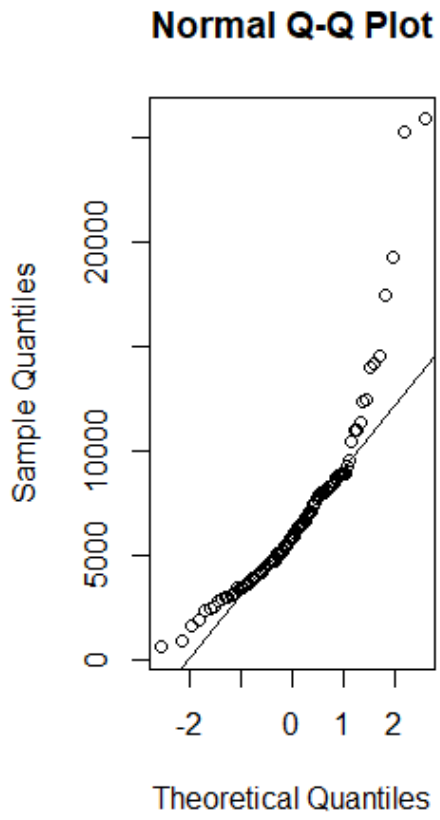




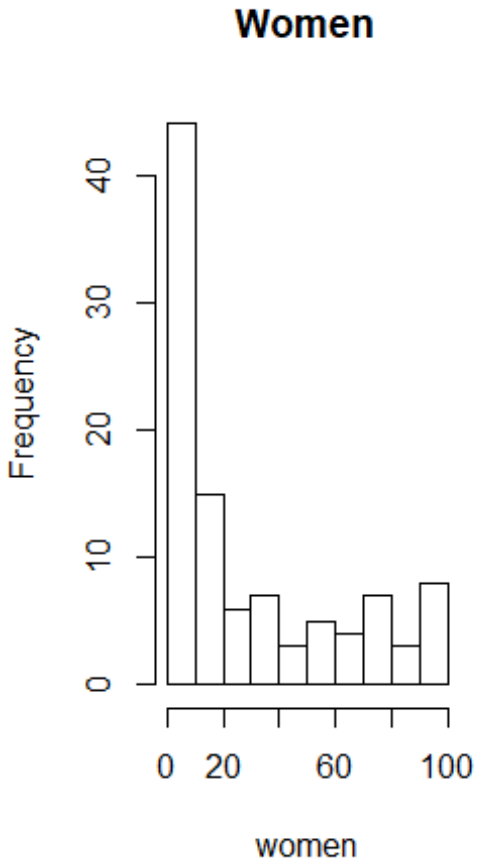
Για το εισόδημα:

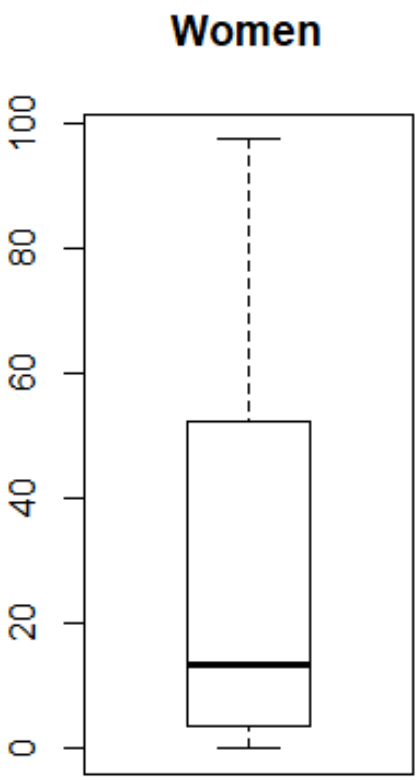


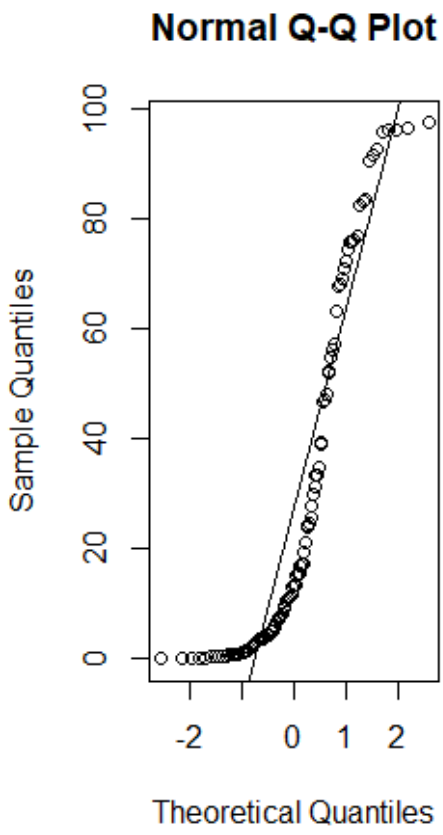




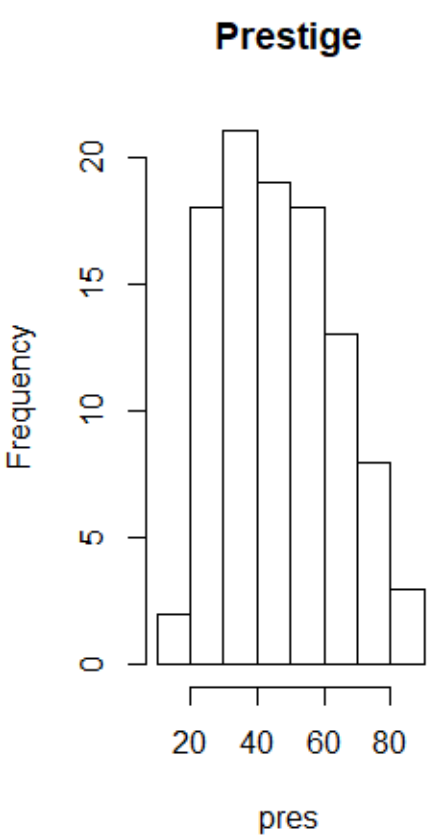
Για τις γυναίκες επαγγελματίες:

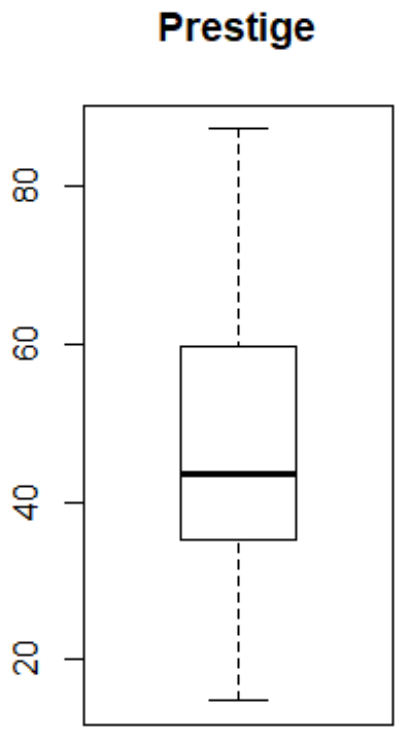


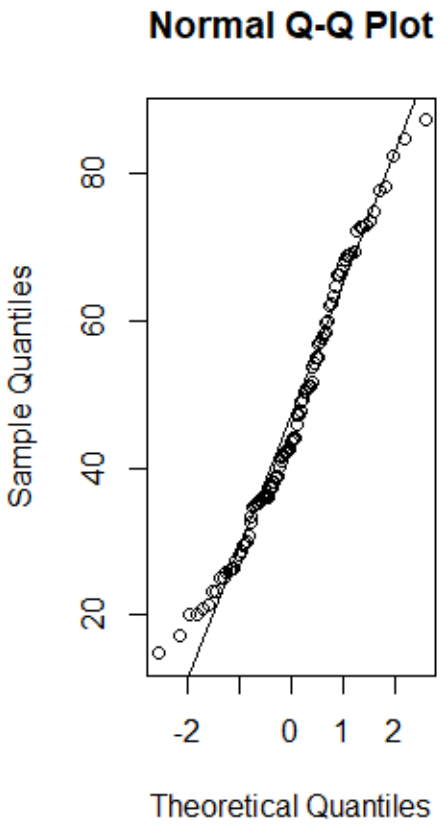




Για την κοινωνικοοικονομική θέση:







Προχωρούμε στο $Shapiro - Wilk \quad Test$ με μηδενική υπόθεση ότι ακολουθείται η κανονική κατανομή και έχουμε τα εξής:
Για εκπαίδευση: $W = 0.94958, \quad p - value = 0.0006773$
Για εισόδημα: $W = 0.81505, \quad p - value = 5.634e - 10$
Για τις γυναίκες επαγγελματίες: $W = 0.81579, \quad p - value = 5.957e - 10$
Για την κοινωνικοοικονομική θέση: $W = 0.97198, \quad p - value = 0.02875.$
Αυτό σημαίνει ότι δεν απορρίπτουμε τη μηδενική υπόθεση μόνο για τη μεταβλητή *prestige* σε επίπεδο στατιστικής σημαντικότητας 1%.
Άρα βρίσκουμε για τη μεταβλητή αυτή της μέση τιμή (46.83333), τη διασπορά (295.9943) και την τυπική απόκλιση (17.20449).
Για τις μεταβλητές που δεν ακολουθούν κανονική κατανομή μπορούμε να βρούμε το μέτρο θέσης *median* και τα ποσοστιμόρια *quantiles*. Παίρνουμε λοιπόν:

Για εκπαίδευση: $median = 10.54, quantiles =$

0%	25%	50%	75%	100%
6.3800	8.4450	10.5400	12.6475	15.9700

Για εισόδημα: $median = 5930.5, quantiles =$

0%	25%	50%	75%	100%
611.00	4106.00	5930.50	8187.25	25879.00

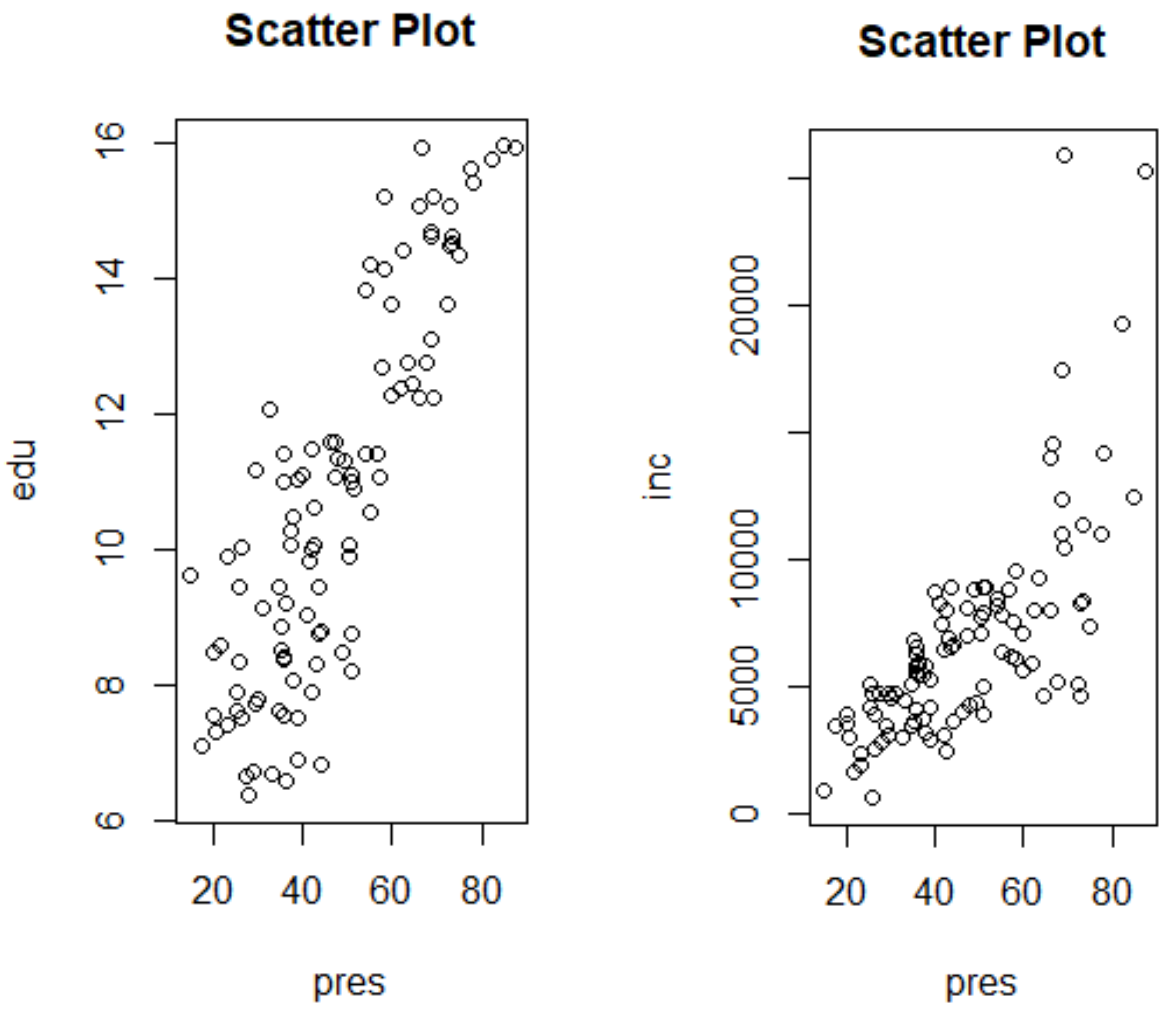
Για τις γυναίκες επαγγελματίες: $median = 13.6, quantiles =$

0%	25%	50%	75%	100%
0	3.5925	13.6	52.2025	97.5100

Παρατηρούμε λοιπόν ότι η μέση μόρφωση είναι 10.54 έτη, με την ελάχιστη στα 6.38 έτη και τη μέγιστη στα 15.97 έτη, το μέσο εισόδημα είναι 5930.5 δολάρια, με το ελάχιστο στα 611 και το μέγιστο στα 25879 δολάρια, ενώ το μέσο ποσοστό των εν ενεργεία γυναικών επαγγελματιών είναι 13.6%, με το ελάχιστο στο 0% και το μέγιστο στο 97.51%.
Το επάγγελμα με το μικρότερο εισόδημα είναι οι *babysitters*, ενώ με το υψηλότερο είναι οι *general managers*.
Το επάγγελμα με το λιγότερο *prestige* είναι τα *newsboys*, ενώ με το μεγαλύτερο είναι οι *physicians*.

2ο Ερώτημα:

α) Παρουσιάζουμε τα δύο σημειογράμματα μεταξύ των μεταβλητών $prestige - education$ και $prestige - income$:



Παρατηρούμε πως είναι γραμμική η σχέση εκπαίδευσης με κοινωνικοοικονομικής θέσης και όσο αυξάνεται η μία μεταβλητή, αυξανεται και η άλλη. Από το σημειόγραμμα μεταξύ εισοδήματος και κοινωνικοοικονομικής θέσης, βλέπουμε πως ενώ μέχρι ένα σημείο υπάρχει έντονη γραμμική σχέση, όταν αυξηθεί λίγο ακόμα το *prestige*, το εισόδημα εκτοξεύεται.

β, γ) Χρησιμοποιώντας τη συνάρτηση lm της R, καταλήγουμε στα εξής:

Θεωρώντας ως εξαρτημένη μεταβλητη το *prestige* και ως ανεξάρτητη την *education* έχουμε κατάλοιπα:

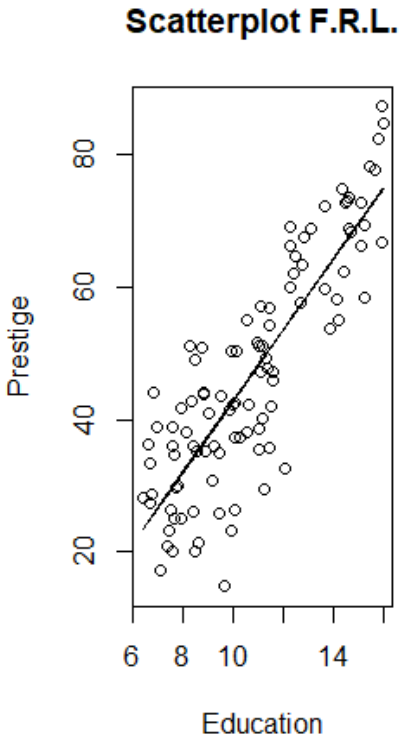
Min $1Q$ $Median$ $3Q$ Max

-26.0397 -6.5228 0.6611 6.7430 18.1636’ οπότε συμπεραίνουμε πως υπάρχει αρκετά σημαντική συμμετρικότητα στα ποσοστημόρια, αλλά όχι στα άκρα των καταλοίπων. Στην συνέχεια, έχουμε βρει τις τιμές των β_0 και β_1 μέσω εκτιμήσεων με το $\hat{\beta}_0$=-10.732 και το $\hat{\beta}_1$=5.361 με αντίστοιχη τυπική απόκλιση 3.677 και 0.332. Κάνουμε έλεγχο στατιστικής σημαντικότητας και προκύπτει ότι και τα δύο είναι σημαντικά για το δείγμα μας, το πρώτο μέχρι 1 τοις χιλίοις και το δεύτερο μέχρι το 0. Οπότε καταλαβαίνουμε ότι η εκπαίδευση μπορεί καλύτερα να προβλέψει την κοινωνικοοικονομική θέση από το ανάποδο. Παραθέτουμε τις τιμές των καταλοίπων και των $\hat{Y}$:

1	2	3	4	5
9.25087492	14.10762099	5.67357335	6.31075828	5.85594955
6	7	8	9	10
4.48785426	2.43633701	6.06002981	5.99203732	1.04873199
11	12	13	14	15
6.31070689	4.79318588	-9.60895705	-4.47909246	8.64977776
16	17	18	19	20
-10.34609058	8.49094016	-7.02443792	-12.56057709	5.79925487
21	22	23	24	25
9.71876461	-2.68317272	-4.01005421	12.37237339	-8.02040906
26	27	28	29	30
0.27347055	8.63544545	-5.02831259	9.81682728	-1.50696831
31	32	33	34	35
9.66635579	8.47984794	0.19522601	3.50354073	-5.40059093
36	37	38	39	40
-8.96450316	-0.55315394	-4.00775709	-2.46758905	-7.52726682
41	42	43	44	45
-21.38102979	-9.75210818	-2.59531071	-6.05205678	-7.51084298
46	47	48	49	50
-19.90984861	2.16541283	-14.84285049	-12.63767307	-0.51905490
51	52	53	54	55
-8.73458717	-16.64483922	-26.03966180	-19.20153390	-4.15419971
56	57	58	59	60
-1.62015206	2.70150060	3.46446986	3.73758837	-1.06121167
61	62	63	64	65
-14.63547874	8.96750436	-14.08192136	-7.76325180	-10.08385870
66	67	68	69	70
-9.80347123	18.16357829	-13.87156651	-1.57261228	9.21096388
71	72	73	74	75
-5.02512389	4.57487611	-5.84573079	8.16770995	3.39966606
76	77	78	79	80
-0.85927433	7.70264916	1.60060903	10.07383034	1.43978270
81	82	83	84	85
7.36347549	14.57069305	-7.23144988	4.72958205	5.40887235
86	87	88	89	90
6.88711689	2.27492751	3.11603851	7.69846610	17.65834947
91	92	93	94	95
12.53470807	11.55018895	-1.23647311	8.97587047	-3.08181857
96	97	98	99	100
11.05401221	14.11813004	5.99652877	-6.67977844	-8.03856464
101	102			
-0.67679534	0.09647737			

```
       1        2        3        4        5        6        7        8
59.54913 54.99238 57.72643 50.48924 67.64405 73.11215 70.16366 72.03997
       9       10       11       12       13       14       15       16
67.10796 67.75127 55.68929 55.20681 63.40896 66.67909 66.25022 65.44609
      17       18       19       20       21       22       23       24
73.80906 65.12444 70.86058 67.00075 74.88124 62.28317 70.11005 74.82763
      25       26       27       28       29       30       31       32
74.72041 68.12653 56.06455 39.92831 62.28317 70.80697 57.83364 48.72015
      33       34       35       36       37       38       39       40
57.40477 50.59646 51.40059 50.86450 49.95315 46.30776 50.16759 38.42727
      41       42       43       44       45       46       47       48
54.08103 48.45211 38.69531 43.25206 45.61084 49.30985 48.93459 50.54285
      49       50       51       52       53       54       55       56
48.23767 42.01905 48.93459 43.14484 40.83966 42.50153 51.45420 48.72015
      57       58       59       60       61       62       63       64
48.39850 40.03553 47.86241 30.76121 34.83548 45.93250 39.98192 28.56325
      65       66       67       68       69       70       71       72
27.38386 29.90347 25.93642 35.37157 36.87261 29.68904 30.22512 30.22512
      73       74       75       76       77       78       79       80
29.04573 25.13229 25.40033 43.35927 36.49735 34.29939 31.72617 34.46022
      81       82       83       84       85       86       87       88
36.33652 36.22931 44.43145 23.47042 32.69113 43.41288 25.02507 37.78396
      89       90       91       92       93       94       95       96
42.50153 33.44165 26.36529 24.64981 31.13647 33.92413 29.58182 55.04599
      97       98       99      100      101      102
34.78187 29.90347 31.77978 34.13856 42.87680 35.10352
```

Χρησιμοποιώντας τα αποτελέσματα της συνάρτησης lm, κάνουμε πάλι σημειόγραμμα και παρατηρούμε ότι η γραμμή είναι ίδια με πριν, συμμετρικά, αφού άλλαξαν θέση οι μεταβλητές:



Όμοια ακριβώς λειτουργούμε για το απλό γραμμικό μοντέλο των $prestige - income$.
όδημα εκτοξεύεται.
Χρησιμοποιώντας τη συνάρτηση lm της R, καταλήγουμε στα εξής:
Θεωρώντας ως εξαρτημένη μεταβλητη το $prestige$ και ως ανεξάρτητη το $income$ έχουμε κατάλοιπα:

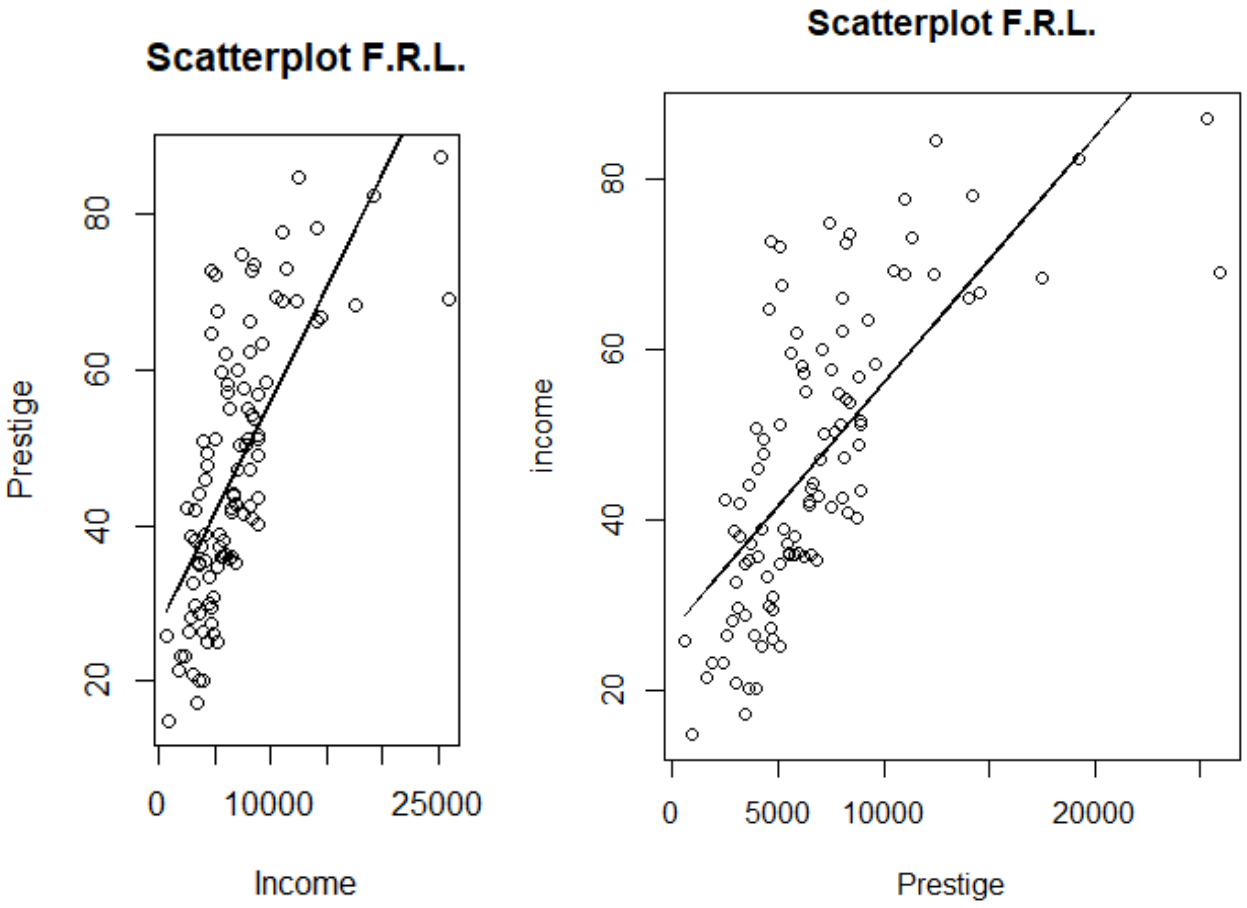
$$\begin{matrix} Min & 1Q & Median & 3Q & Max \\ -33.007 & -8.378 & -2.378 & 8.432 & 32.084 \end{matrix}$$

, οπότε συμπεραίνουμε πως αφού η διάμεσος απέχει πολύ από το 0 δεν υπάρχει η συμμετρικότητα που είχαμε πριν, αλλά τα ποσοστημόρια και τα άκρα εμφανίζουν μεταξύ τους σημαντική συμμετρικότητα. Στην συνέχεια, έχουμε βρει τις τιμές των β_0 και β_1 μέσω εκτιμήσεων με το $\hat{\beta}_0$=27.141176368 και το $\hat{\beta}_1$=0.002896799 με αντίστοιχη τυπική απόκλιση $2.268e{+}00$ και $2.833e$-04. Κάνουμε έλεγχο στατιστικής σημαντικότητας και προκύπτει ότι και τα δύο είναι σημαντικά για το δείγμα μας, με κοινή τιμή $2e$-16, δηλαδή κοντά στο 0. Οπότε καταλαβαίνουμε ότι η μία μεταβλητή προβλέπει την άλλη στο ίδιο επίπεδο. Παραθέτουμε τις τιμές των καταλοίπων και των $\hat{Y}$:

1	2	3	4	5	6
5.8804567	-33.0074429	9.4025982	3.9786987	22.0170199	18.5071285
7	8	9	10	11	12
21.5370558	9.9314566	13.0019391	9.7274060	17.7619148	12.4103181
13	14	15	16	17	18
2.2532904	11.7424869	26.3080256	9.6047039	-0.6422194	13.2535869
19	20	21	22	23	24
3.3698289	32.0844226	21.3067696	16.0977017	15.6859389	-13.2533705
25	26	27	28	29	30
-2.6127791	-9.4293688	24.1929921	-2.3365216	30.2083221	11.9394144
31	32	33	34	35	36
25.3534038	12.1073590	8.5532281	3.1876894	7.1673421	5.6396997
37	38	39	40	41	42
9.6635407	8.0674592	8.0156831	-10.0328374	-3.1779228	3.1552092
43	44	45	46	47	48
-7.0054368	-0.7723086	1.8020414	-11.4749014	9.3241941	-9.5722426
49	50	51	52	53	54
-3.3456331	-7.3150280	-12.3750733	-8.1554735	-15.0004380	-10.7065905
55	56	57	58	59	60
-3.3950507	-0.2955964	0.9118892	-9.4082053	-1.2966181	-6.4676027
61	62	63	64	65	66
-18.3255972	4.9639107	-3.0111207	-15.0315740	-19.8988632	-17.4175111
67	68	69	70	71	72
6.4057841	-10.4382758	-11.7132189	-0.4048362	-16.8133435	-7.2133435
73	74	75	76	77	78
-9.4161269	-6.7116552	-8.4365216	-7.9401323	-2.3091758	-10.2586631
79	80	81	82	83	84
-4.1037448	-8.0744765	-2.4818375	12.2396412	-5.7258352	-7.1883637
85	86	87	88	89	90
-5.8281277	0.8071210	-13.4445454	-10.3309585	2.3553997	-1.7647533
91	92	93	94	95	96
-3.5913153	-8.2032028	-10.4187159	-4.3102012	-11.9676612	-1.6890627
97	98	99	100	101	102
-3.8633653	-7.3531735	-14.2772562	-14.8096630	-3.6602928	-2.4188991

1	2	3	4	5	6	7
62.91954	102.10744	53.99740	52.82130	51.48298	59.09287	51.06294
8	9	10	11	12	13	14
68.16854	60.09806	59.07259	44.23809	47.58968	51.54671	50.45751
15	16	17	18	19	20	21
48.59197	45.49530	82.94222	44.84641	54.93017	40.71558	63.29323
22	23	24	25	26	27	28
43.50230	50.41406	100.45337	69.31278	77.82937	40.50701	37.23652
29	30	31	32	33	34	35
41.89168	57.36059	42.14660	45.09264	49.04677	50.91231	38.83266
36	37	38	39	40	41	42
36.26030	39.73646	34.23254	39.68432	40.93284	35.87792	35.54479
43	44	45	46	47	48	49
43.10544	37.97231	36.29796	40.87490	41.77581	45.27224	38.94563
50	51	52	53	54	55	56
48.81503	52.57507	34.65547	29.80044	34.00659	50.69505	47.39560
57	58	59	60	61	62	63
50.18811	52.90821	52.89662	36.16760	38.52560	49.93609	28.91112
64	65	66	67	68	69	70
35.83157	37.19886	37.51751	37.69422	31.93828	47.01322	39.30484
71	72	73	74	75	76	77
42.01334	42.01334	32.61613	40.01166	37.23652	50.44013	46.50918
78	79	80	81	82	83	84
46.15866	45.90374	43.97448	46.18184	38.56036	42.92584	35.38836
85	86	87	88	89	90	91
43.92813	49.49288	40.74455	51.23096	47.84460	52.86475	42.49132
92	93	94	95	96	97	98
44.40320	40.31872	47.21020	38.46766	67.78906	52.76337	43.25317
99	100	101	102			
39.37726	40.90966	45.86029	37.61890			

Χρησιμοποιώντας τα αποτελέσματα της συνάρτησης lm, κάνουμε πάλι σημειόγραμμα και παρατηρούμε ότι η γραμμή είναι ίδια με πριν, σχετικά συμμετρικά, αφού άλλαξαν θέση οι μεταβλητές:



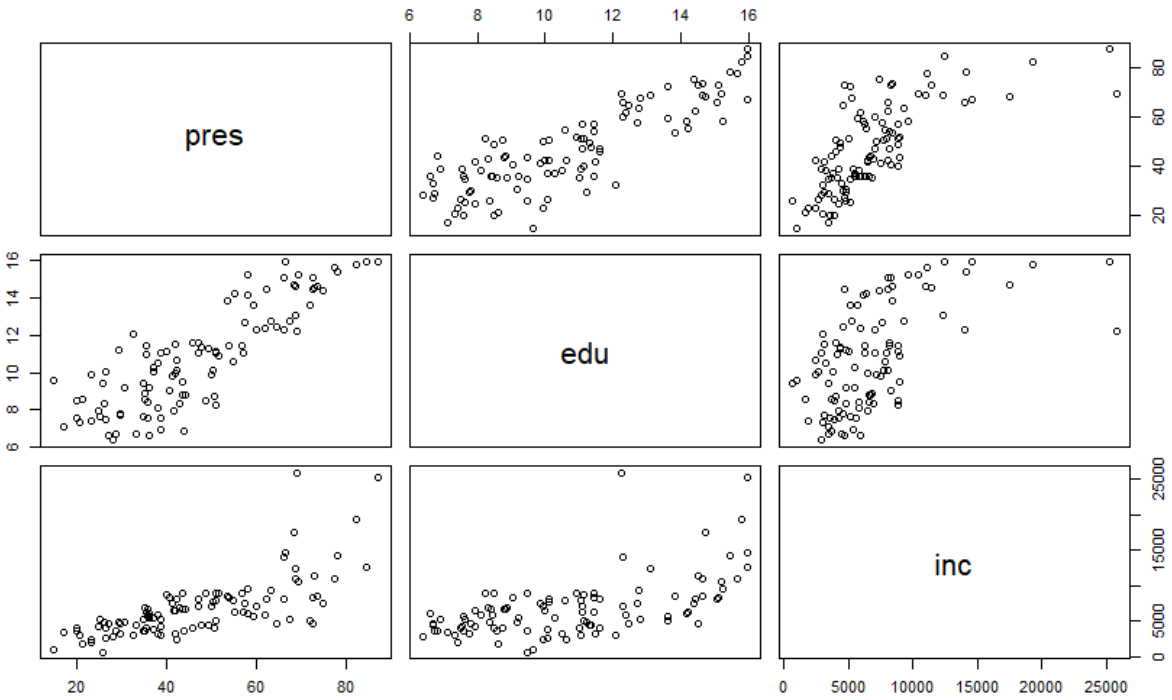
δ) Για να κάνουμε έλεγχο υποθέσεων χρειάστηκε να εκτιμήσουμε τα β_0, β_1 και το Υ. Με αυτόν τον τρόπο υπολογίζουμε τα κατάλοιπα, τη διακύμανση και στη συνέχεια τη διακύμανση των β_0, β_1. Έχοντας ως υποθέσεις τις $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 > 0$ βρίσκουμε την παρατηρούμενη τιμή της ελεγχοσυνάρτησης $t^* = 16.1478$. Για επίπεδο στατιστικής σημαντικότητας α=0.05 με 100 βαθμούς ελευθερίας έχουμε ότι το ποσοστημόριο είναι 1.660234, πολύ μικρότερο από την παρατηρούμενη τιμή, άρα απορρίπτουμε την H_0 και δεχόμαστε την εναλλακτική υπόθεση.

ε) Μέσω της συνάρτησης lm βρίσκουμε τα διαστήματα εμπιστοσύνης σε επίπεδο στατιστικής σημαντικότητας 10% για τους συντελεστές του μοντέλου $prestige - education$ να κυμαίνονται από -16.83% έως -4.63% για το β_0 και από 4.8% έως 5.91% για το β_1. Μέσω της συνάρτησης lm βρίσκουμε ομοίως τα διαστήματα εμπιστοσύνης σε

επίπεδο στατιστικής σημαντικότητας 10% για τους συντελεστές του μοντέλου $prestige - income$ να κυμαίνονται από 23.37% έως 30.9% για το β_0 και από 0.0024% έως 0.003367% για το β_1.
στ) Υπολογίζουμε μέσω του *summary* τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού για τα δύο μοντέλα κι έχουμε ότι για το $prestige - education$ $R^2 = 0.7228$, $R^2_{adj} = 0.72$ και για το $prestige - income$ $R^2 = 0.5111$, $R^2_{adj} = 0.5062$. Στη γραμμική παλινδρόμηση του $prestige - education$ το R^2 είναι αρκετά μεγάλο ώστε να έχει πολλή πληροφορία, ενώ στο $prestige - income$ είναι 0.5 περίπου ο συντελεστής προσδιορισμού. Αυτό σημαίνει ότι η μισή πληροφορία βρίσκεται σε σφάλματα οπότε χρειάζονται περισσότερα στοιχεία για καλό αποτέλεσμα, δηλαδή καλή προσαρμογή του μοντέλου.
ζ) Με βάση τα παραπάνω βρίσκουμε σημειακή εκτίμηση για την τιμή 42.87 σε διάστημα 41.32602 με 44.42757. Αντίστοιχα, το ατομικό διάστημα έχει ίδια πρόβλεψη, αλλά σε διάστημα 27.68385 με 58.06974.

3ο Ερώτημα:

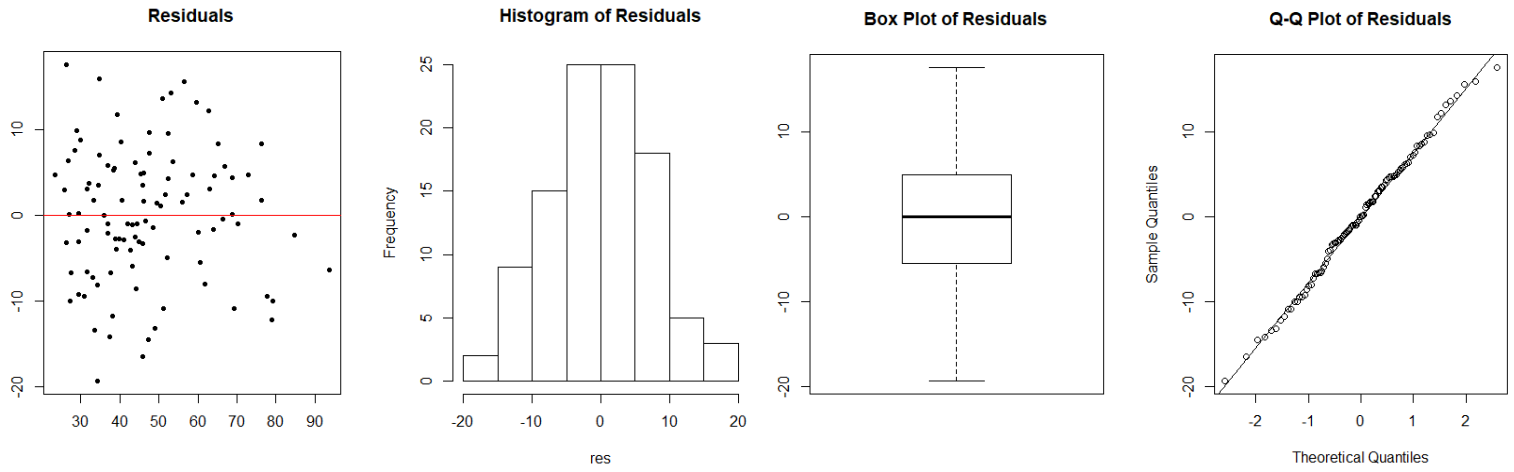
α) Θέλουμε να οπτικοποιήσουμε τις σχέσεις των μεταβλητών που επιλέξαμε, οπότε κάνουμε το εξής σημειόγραμμα:



Δεδομένου ότι πήραμε ως εξαρτημένη μεταβλητή την κοινωνικοοικονομική θέση, εξετάζουμε την πρώτη γραμμή των σημειογραμμάτων. Παρατηρούμε θετική γραμμική σχέση πιο έντονη με την ανεξάρτητη μεταβλητή της εκπαίδευσης. Τα διαστήματα εμπιστοσύνης που εξάγουμε είναι:

```
                  2.5 %       97.5 %
(Intercept) -1.323493e+01 -0.460629799
edu          3.445127e+00  4.829761535
inc          9.162805e-04  0.001806051
```

Όσον αφορά τις παραμέτρους της παλινδρόμησης, παρατηρούμε από τα $p-values$, τα οποία βγήκαν $< 2e-16$ και $2.36e$-08 για την εκπαίδευση και το εισόδημα αντίστοιχα, ότι είναι στατιστικά σημαντικές και οι δύο μεταβλητές για το μοντέλο μας.
β) Μέσω της συνάρτησης lm υπολογίζουμε τον συντελεστή προσδιορισμού: 0.798 και τον προσαρμοσμένο: 0.7939. Αυτό σημαίνει ότι εξηγείται ένα 80% της μεταβλητότητας της κοινωνικοοικονομικής θέσης.
γ) Οι αρχικές υποθέσεις ελέγχονται μέσω του $F-test$. Κάνουμε τον έλεγχο για ολόκληρο το μοντέλο και καταλήγουμε σε τέτοιο $p-value$ που δείχνει ότι το μοντέλο μας έχει νόημα όπως είναι. Κάνουμε τώρα γραφικό έλεγχο των καταλοίπων.

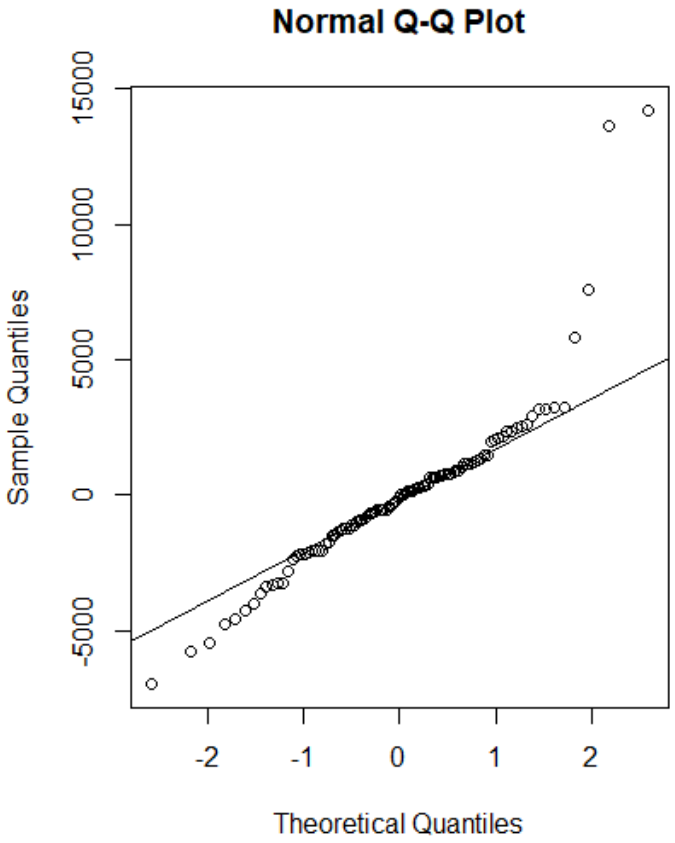


Όπως παρατηρούμε από τα παραπάνω, το ιστόγραμμα λόγω της ομοιομορφίας του στο κέντρο μας παραπέμπει σε κανονική κατανομή, αλλά για επιβεβαίωση προχωρούμε στο $boxplot$ της παλινδρόμησης, όπου το κουτί του διαγράμματος σχεδόν ισαπέχει από τις άκρες και χωρίς ακραίες παρατηρήσεις. Σχεδιάζουμε στο τέλος το $q-qplot$ που δείχνει κανονική κατανομή με ελάχιστες παρατηρήσεις να μην ακουμπούν τη γραμμή της κανονικής. Ολοκληρώνουμε τον έλεγχο με ένα $Shapirotest$, που δίνει $p-value$=0.9371. Αν θεωρήσουμε H_0 την υπόθεση ότι το δείγμα ακολουθεί κανονική κατανομή και H_1 την εναλλακτική υπόθεση ότι το δείγμα δεν ακολουθεί κανονική κατανομή, έχουμε ότι αν απορρίψουμε την H_0 θα έχουμε σφάλμα της τάξης του 93%, οπότε δεν την απορρίπτουμε.

δ) Συγκρίνοντας τις τιμές των R^2 και R^2_{adj} από τα μοντέλα έχουμε ότι στις απλές παλινδρομήσεις εξηγείται μικρότερο ποσό πληροφορίας σε σχέση με το πολλαπλό μοντέλο. Η διαφορά δεν είναι μεγάλη αφού για $prestige$ και εκπαίδευση R^2=0.7228 και για $prestige$ και εισόδημα R^2=0.5111, ενώ στην πολλαπλή παλινδρόμηση R^2=0.798. Επομένως, θα προτιμήσουμε την ανάλυση με πολλαπλή παλινδρόμηση.
ε) Έχουμε βρει στο Ερώτημα 2 ότι σημειακή εκτίμηση 42.87 σε διάστημα 41.32602 με 44.42757. Αντίστοιχα, το ατομικό διάστημα έχει ίδια πρόβλεψη, αλλά σε διάστημα 27.68385 με 58.06974. Τώρα, βρίσκουμε σημειακή εκτίμηση 39.97133 σε διάστημα 38.42149 με 41.52116. Αντίστοιχα, το ατομικό διάστημα έχει ίδια πρόβλεψη, αλλά σε διάστημα 26.91113 με 53.03152. Βλέπουμε πάλι ότι η ατομική πρόβλεψη είναι πιο ευρεία για να εξασφαλίσουν επιτυχή πρόβλεψη.
στ) Για να κάνουμε έλεγχο υποθέσεων χρειάστηκε να εκτιμήσουμε τα β και το Υ. Με αυτόν τον τρόπο υπολογίζουμε τα κατάλοιπα, τη διακύμανση και στη συνέχεια τη διακύμανση των β. Έχοντας ως υποθέσεις τις $H_0 : \beta_{edu} = 4 \quad vs \quad H_1 : \beta_{edu} > 4$ βρίσκουμε την παρατηρούμενη τιμή της ελεγχοσυνάρτησης $t^* = -3.280814e + 12$. Για επίπεδο στατιστικής σημαντικότητας α=0.05 με 100 βαθμούς ελευθερίας έχουμε ότι το ποσοστημόριο είναι 1.660715, πολύ μικρότερο από την παρατηρούμενη τιμή, άρα δεν μπορούμε να απορρίψουμε την H_0.

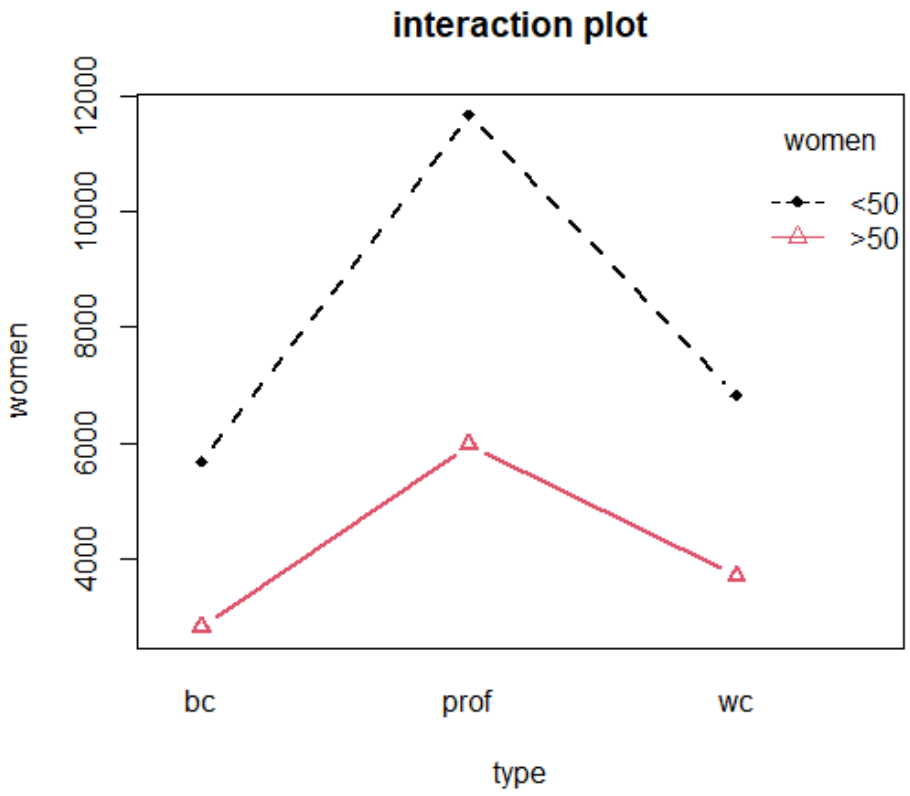
4ο Ερώτημα:

α, β) Υποθέσεις του μοντέλου $ANOVA$: ι) η κατανομή των παρατηρήσεων σε κάθε επίπεδο είναι κανονική για το επίπεδο i, ιι) η κανονική κατανομή σε κάθε επίπεδο έχει την ίδια διασπορά σ^2. Ισοδύναμα, οι τυχαίοι όροι $\epsilon_{i,j}$ είναι ταυτοτικά κατανεμημένοι για κάθε i, j, ιιι) οι παρατηρήσεις σε κάθε επίπεδο του παράγοντα είναι ανεξάρτητες και ταυτοτικά κατανεμημένες και είναι ανεξάρτητες από τις παρατηρήσεις στα άλλα επίπεδα.
Κάνουμε τις κατάλληλες αλλαγές ώστε η μεταβλητή $women$ να παίρνει τις τιμές που ζητήθηκαν και διαχωρίζουμε εξαρτημένη και ανεξάρτητες μεταβλητές. Φτάνουμε έτσι σε ένα μοντέλο $ANOVA$ 2 παραγόντων, όπου και οι δύο ανεξάρτητες μεταβλητές είναι στατιστικά σημαντικές, αλλά η αλληλεπίδρασή τους όχι, με $p - value$=0.271. Βγάζοντας την αλληλεπίδραση από το μοντέλο, καταλήγουμε πως μπορεί διαφορετικά δρουν οι γυναίκες επαγγελματίες από τις χειρώνακτες μεταξύ τους, αλλά με τον ίδιο τρόπο σε σχέση με το εισόδημα. Έτσι, καταλήγουμε στην τελική μορφή του μοντέλου, που περιέχει τις μεταβλητές $women, type$, αλλά όχι την αλληλεπίδρασή τους.
γ) Θέλουμε να ελέγξουμε τα κατάλοιπα οπότε κάνουμε το $q - qplot$:



Παρατηρούμε ότι στις ουρές χαλάει η κανονικότητα οπότε ελέγχουμε με *Shapirotest* και βρίσκουμε $p - value = 5.38e - 09$, τόσο μικρό που απορρίπτουμε την κανονικότητα και στηριζόμαστε στο μεγάλο δείγμα για την αξιοπιστία του μοντέλου. Προηγουμένως, είχαμε οδηγηθεί σε κατάλοιπα που ακολουθούσαν κανονική κατανομή σε αντίθεση με τώρα.

δ) Κάνουμε το *interaction plot* και βλέπουμε ότι οι δύο μεταβλητές έχουν παρόμοια μορφή, άρα έχουν την ίδια επίδραση περίπου στο μοντέλο. Μάλιστα, αυτός είναι ένας πολύ καλός τρόπος να καταλάβουμε ότι σωστά βγάλαμε τον όρο της αλληλεπίδρασης προηγουμένως από το μοντέλο, καθώς παρατηρούμε μικρή διαφορά:


interaction plot
women
4000
6000
8000
10000
12000
bc
prof
wc
type
women
<50
>50

ΚΩΔΙΚΑΣ R:

```r
data<-read.csv(file=prestige, header=TRUE)
head(prestige)
profs<-prestige$Professions
edu<-prestige$education
inc<-prestige$income
women<-prestige$women
pres<-prestige$prestige
type<-prestige$type
type_freq=table(type)
type_freq
profs_freq=table(profs)
barplot(type_freq,main="Type of Worker", xlab="Number of workers", col="black")
pie(type_freq, main="Type of worker", col=c("grey10", "grey20", "grey30"))
par(mfrow=c(1,2))
hist(edu,main="Education")
hist(inc, main="Income")
hist(women, main="Women")
hist(pres, main="Prestige")
boxplot(edu,main="Education")
boxplot(inc, main="Income")
boxplot(women, main="Women")
boxplot(pres, main="Prestige")
qqnorm(edu)
 qqline(edu)
qqnorm(inc)
qqline(inc)
qqnorm(women)
qqline(women)
qqnorm(pres)
qqline(pres)
shapiro.test(edu)
shapiro.test(inc)
shapiro.test(women)
shapiro.test(pres)
mu=mean(pres)
mu
variance=var(pres)
variance
sd=sd(pres)
sd
median(edu)
quantile(edu)
median(inc)
quantile(inc)
median(women)
quantile(women)
min(inc)
incom<-which.min(inc)
profs[incom]
max(inc)
inco<-which.max(inc)
profs[inco]
min(pres)
prestig<-which.min(pres)
profs[prestig]
max(pres)
```

```
prest<-which.max(pres)
profs[prest]
plot(pres,edu,
     main="Scatter Plot",
     preslab="Prestige",
     edulab="Education")
plot(pres,inc,
     main="Scatter Plot",
     preslab="Prestige",
     inclab="Income")
y<-prestige$pres
x<-prestige$edu
reg<-lm(y~x)
summary(reg)
reg$coefficients
reg$residuals
reg$fitted.values
plot(x,y,
     main="Scatterplot F.R.L.",
     xlab="Education",
     ylab="Prestige")
lines(x,reg$fitted.values)
x<-prestige$inc
reg<-lm(y~x)
summary(reg)
reg$coefficients
reg$residuals
reg$fitted.values
plot(x,y,
     main="Scatterplot F.R.L.",
     xlab="Income",
     ylab="Prestige")
lines(x,reg$fitted.values)
reg1=lm(pres~inc)
reg1$coefficients
reg1$residuals
reg1$fitted.values
z=pres
w=inc
plot(w,z,
     main="Scatterplot F.R.L.",
     xlab= "Prestige",
     ylab="income")
lines(w,reg1$fitted.values)
shapiro.test(edu)
shapiro.test(inc)
shapiro.test(pres)
x<-edu
t<-sum((x-mean(x))^2)
b1<-sum((x-mean(x))*(y-mean(y)))/t
b0<-mean(y)-b1*mean(x)
yhat<-b0+b1*x
ehat<-y-yhat
sigma2hat<-sum(ehat^2)/(100)
sigmahat<-sqrt(sigma2hat)
```

```
sigma2_b0<-sigma2hat*(1/100+mean(x)^2/t)
std_b0=sqrt(sigma2_b0)
sigma2_b1<-sigma2hat/t
std_b1=sqrt(sigma2_b1)
c = 0
t_stat<-(b1-c)/std_b1
t_stat
a<-0.05
qt_a<-qt(1-a, 100)
qt_a
t<-sum((w-mean(w))^2)
b1<-sum((w-mean(w))*(y-mean(y)))/t
b0<-mean(y)-b1*mean(w)
yhat<-b0+b1*w
ehat<-y-yhat
n<-102
sigma2hat<-sum(ehat^2)/(n-2)
sigmahat<-sqrt(sigma2hat)
sigma2_b0<-sigma2hat*(1/n+mean(w)^2/t)
std_b0=sqrt(sigma2_b0)
sigma2_b1<-sigma2hat/t
std_b1=sqrt(sigma2_b1)
c = 0
t_stat<-(b1-c)/std_b1
t_stat
a<-0.05
qt_a<-qt(1-a, n-2)
qt_a
reg=lm(pres~edu)
reg1=lm(pres~inc)
confint(reg, level=0.90)
confint(reg1, level=0.90)
summary(reg)
summary(reg1)
regg=lm( pres ~ edu + inc )
summary(regg)
confint(regg)
futurex<-10
futurew<-4000
future<-data.frame(edu=futurex,  inc=futurew)
predict(reg, newdata=future, level=0.90, interval="confidence")
predict(reg, newdata=future, level=0.90, interval="prediction")
head(prestige)
pres<-prestige$prestige
edu<-prestige$education
inc<-prestige$income
pairs(cbind(pres,edu,inc))
reg <- lm(pres ~ edu + inc)
summary(reg)
confint(reg)
coef=reg$coefficients
res =reg$residuals
fitted=reg$fitted.values
plot(fitted,res,main = "Residuals",xlab = "",ylab = "",
     pch=20)
```

```
lines(x=c(0,100),y=c(0,0),col=("red"))
hist(res,main="Histogram of Residuals")
boxplot(res,main="Box Plot of Residuals")
n=length(profs)
ones=rep(1,n)
X=cbind(ones,edu,inc)
betas=solve(t(X)%*%X)%*%t(X)%*%pres
ehat=y-X%*%betas
qqnorm(res,main="Q-Q Plot of Residuals")
qqline(ehat)
shapiro.test(res)
simple.edu<-lm(pres~edu)
simple.inc<-lm(pres~inc)
summary(simple.edu)
summary(simple.inc)
summary(reg)
res.edu =simple.edu$residuals
res.inc =simple.inc$residuals
res =reg$residuals
edu.avg<-10
inc.avg<-4000
newjob<-data.frame(edu=edu.avg, inc= inc.avg)
predict(reg, newdata=newjob, level=0.90, interval="confidence")
predict(reg, newdata=newjob, level=0.90, interval="prediction")
betas<-solve(t(X)%*%X)%*%t(X)%*%y
betas
yhat<-X%*%betas
ehat<-y-X%*%betas
ehat
sigma2.hat<-t(ehat)%*%ehat/(n-2)
sigma.hat<-sqrt(sigma2.hat)
var.betas<-as.vector(sigma2.hat)*solve(t(X)%*%X)
std.betas<-sqrt(diag(var.betas))
var.betas
std.betas
betas
c<-4
t_stat<-(betas[2]-c)/std.betas[2]
t_stat
a<- 0.05
qt_a<-qt(1-a, n-5)
qt_a
install.packages("gplots")
library(gplots)
head(prestige)
women<-prestige$women
women[women<50]<-0
women[women >=50]<-1
y<-prestige$income
n<-length(y)
y
n
type<-prestige$type
women
type
women<-factor(women)
levels(women)<- c("<50",">50")
two.way.aov<-aov(y~ women*type)
summary(two.way.aov)
two.way.aov<-aov(y~ women+type)
summary(two.way.aov)
resid.twoway<-two.way.aov$residuals
qqnorm(resid.twoway)
qqline(resid.twoway)
shapiro.test(resid.twoway)
interaction.plot(type,women,y,type="b",col=c(1:3),leg.bty="n",lwd=2,
                 pch=c(18,24),xlab="type",ylab="women",main="interaction plot")
```